Proposal Regarding An Enterprise Data Repository

18 January 2001

Introduction

Last summer the DBA staff began looking at solutions to three problems, two of which came out of Data Council meetings and a third that came from the Pre-Integration Group. The basic problem was developing a metadata Enterprise Data Repository that:

- Supported the Installation Core Business Model (Navy) hierarchy
- Provided for data migration (source-to-target mapping)
- Provided for a standardized naming convention

Enterprise Data Repository

For the first two items, we looked at a 1998 Data Council presentation titled, "Managing SSIIS Legacy Data Migration in an ICBM World". That presentation provided a high-level logical data model that would answer questions like:

- What enterprise data is available? Required?
- What legacy data is available? Required?
- What applications depend on which data sources?
- What data elements are spelled the same?
- How do their meanings compare?
- What required data sources are not currently available?
- How many available data sources exist for a required data element?

The suggested "Next Steps" in this presentation was to:

- Develop and implement the logical data model in an Oracle database (implied).
- Acquire or develop a repository tool. If you remember, the Data Council has considered some data conversion/data repository tools in the past, but for one reason or another did not pursue their investigation/evaluation.
- Develop and test baseline procedures and tool sets with ongoing projects.

Building the Repository

The first step was to build the metadata repository using the logical data model mentioned above as a guide. The repository was built in Oracle 8.1.6 and was designed to hold:

- The ICBM hierarchy, from Core Business Area down to the Sub-Functional level.
- The Enterprise Application and sub-application information.
- The Legacy Application and sub-application information.
- The Enterprise Element information, from instance name down to the column level.
- The Legacy Element information, from instance name down to the column level.
- Point-of-contact and data steward/custodian information.
- Data conversion information.
- Standardized abbreviated names.

Repository Tools

The staff then began looking at what we could use to develop a repository tool. Two products - Oracle Designer and Platinum ERWin - were considered and dismissed because they require a thick client and are not consistently up-to-date. On this latter point, what I mean is that table and relationship changes made in the database are not automatically carried over to the data modeling tools, that it's a time-consuming, manual process.

That's when we considered Oracle's internal data dictionary features. These tables are always up-to-date and accessible through SQLPLUS or over the intranet. In addition, they could contain much of the metadata we require if the table and column comment tables were populated.

Table and Column Definitions

With this in mind, the DBA staff began adding definitions to the MMDB, BASICS, GRASP, ETAC, METERS, and MWR Drillhall tables and columns, and we asked AMS (Penny, Ezi) to do the same to their applications. The format we asked everyone to follow is shown below. Note that we asked everyone to fully define (or spell out) the attribute name as the first part of the column comments.

COMMENT ON TABLE BASICS.BADGE IS 'This table contains the information required by the BASICS system for a badge to be issued.' ;

COMMENT ON COLUMN BASICS.BADGE.CY_CD_FK IS '**COUNTRY CODE FOREIGN KEY.** This field provides a foreign key relationship to the BASICS.COUNTRIES table, identifying the country of residence for this badge holder.' ;

Using four data dictionary tables, static reference tables, and several SQL scripts,

- DBA_TABLES
- DBA_TAB_COLUMNS
- DBA_TAB_COMMENTS
- DBA_COL_COMMENTS

The DBA staff could then populate the Enterprise-related tables:

- ICBM
- Enterprise Applications
- Enterprise Elements

These tables contain the metadata for each data element including abbreviated name, full name, data type, size, precision, scale, mandatory/optional, definitions, notes, and comments.

Source-to-Target Mapping

To accommodate the source-to-target mapping requirement, a many-to-many relationship and a conversion rule table were added to the repository. At this time, the mapping process is expected to be a manual process, though some simple automated processes could be provided.

Standard Naming Conventions

The third problem we wrestled with was standardizing names for new applications coming into SSIIS. Pre-Integration developers wanted to modify their table and column names to SSIIS standards before the migration took place in order to minimize the impact to their applications. The problem was that at the time the request was made, there wasn't a standard naming and abbreviation convention.

The Importance of Standardized Names

For insight into this issue we looked at a 1998 Data Council presentation titled, "A Strategy for Data Sharing", which said in part that "data from disparate sources can be shared, transferred, stored, and analyzed effectively only when it has the same name, meaning (definition), and physical characteristics."

The report went on to say that for data to retain its value, it "must be projected consistently across a wide variety of functional areas and organizations, but this cannot be done unless standard names, definitions, and codes are used consistently across different organizations and systems."

Being consistent in our names, definitions, and codes is something that everyone, including the DBA staff, has failed to do. For example, we have "people" type information in at least three places:

- Personnel data in Military Manpower Data Basde
- Point-of-contact information in Protocol
- Person/security information in our Access Management Approach.

The problem gets worse when one considers data element names. For example, in the Badging and Security Information Control System database, there are multiple abbreviations for:

- License – LIC, LICNS
- Issuance – ISN, ISU
- Country – CY, CNTRY

And we haven't even considered the problems with synonyms (where two data elements have different names but the same meanings) and homonyms (where two data elements have the same name but different meanings).

Benefits of Data Standardization

Data standardization takes time and effort, but we believe it provides enormous benefits including:

- Reduced redundancy, inconsistency, and maintenance costs
- Improved communication/coordination between decision makers
- Improved system interoperability

Implementation Approach

We recommend that SSIIS data standardization should be approached at two levels:

- At the global level
  - Ensure that one source for the data is used across the enterprise. For example, this means that applications requiring:
    - Organization information use the FMM.ORGANIZATION table;
    - Officer-rank information use the MMDB.OFFICER_RANK_PAY_REF table;
    - Badging information use the BASICS.BADGE table.
  - Ensure that table and column names, definitions, and codes are projected consistently across the enterprise.
  - Don't create localized tables.
  - Establish database links, userids, and synonyms are created as needed to access data in other databases, instances, or machines. (This process is already in place and used daily by SSIIS applications.)
  - Submit our applications to DISA for approval and inclusion into the DDDS.
- At the atomic level
  - Establish the Enterprise Data Repository and mandate its use by all parties developing and maintaining SSIIS applications.
  - Use DISA's table and column naming and definition guidelines.
  - Use a consistent abbreviation methodology.

DISA Entity and Data Element Naming and Definition Guidelines

The DOD Data Standardization Procedures (DoD 8320.1-M-1) provide for the addition of new entities and attributes into the Data Dictionary System (DDS). These procedures summarized below include excellent naming and definition guidelines for both entities and attributes.

Entity Naming and Definition Guidelines

The guidelines for naming entities specify:

- The entity name shall be a singular noun or noun phrase, and include only alphabetic characters and hyphens.
- The entity name shall not contain class word names except under special circumstances, abbreviations, acronyms, organization names, articles (a, an, the, etc.), or prepositions (at, by, for, etc.).

The guidelines for defining entities specifies:

- The entity definition should define what the entity is, not how, where, or when the entity is used, or who uses it; add meaning to the name; be concise, precise, and unambiguous; and avoid circular reasoning.
- The entity definition should not contain examples or infinitives ("This entity defines").

Data Element Naming and Definition Guidelines

The guidelines for naming data elements specify:

- The data element name shall be composed of:
  - The entity name (customer)
  - A property modifier (delivery)
  - A class word modifier (month)
  - A class word (code)
- The data element name shall be a singular noun phrase, and include only alphabetic characters and hyphens.
- The data element name shall not contain abbreviations, acronyms, organization names, articles (a, an, the, etc.), prepositions (at, by, for, etc.), or possessive forms of a word.

The guidelines for defining data elements specifies:

- The data element definition should define what the data is, not how, where, or when the data is used, or who uses it; be comprised of simple sentences; represent a characteristic of its associated entity; spell out any acronyms and abbreviations; be concise, precise, and unambiguous; and avoid circular reasoning.
- The data element definition should not contain examples or physical characteristics of the data element, or infinitives ("This data element defines…").

Abbreviation Proposal

The DBA staff also suggested guidelines to be used in abbreviating new data elements migrating into Shore Station managed databases. The goal was to provide a consistent means of abbreviating data element names to a reasonable length without sacrificing readability.

We recognized from the outset that there were at least five sources that we would use to abbreviate new data elements:

- Defense Data Dictionary System (DDDS)
- Established Navy standards.
- Non-local Navy databases such as the Naval Facility Assets Database (NFADB)
- Local Shore Station databases
- Commonly used abbreviations

Defense Data Dictionary System

The data dictionary would use abbreviations approved by and used in the DDDS. For example, the class words required by the DDDS have designated abbreviations that are to be used in naming any data element. These are:

Amount – AM	Name – NM
Angle – AN	Quantity – QY
Area – AR	Rate – RT
Code – CD	Temperature – TP
Coordinate – CN	Text – TX
Date – DT	Time – TM
Dimension – DM	Volume – VL
Identifier – ID	Weight – WT
Mass – MS	

Navy Standard Abbreviations

The data dictionary would also use established Navy standard abbreviations where such abbreviations were known:

- ACCT – Account
- ADMIN – Administration
- EMP – Employee
- ORG – Organization

Non-Local Naval Abbreviations

The data dictionary would use data element abbreviations from other Naval databases such as the NFADB. For example:

- APR – Acquisition Property Record
- BOD – Beneficial Occupancy Date
- CIP – Capital Improvement Plan
- CRDN – Cost Reference Document Number

Local Shore Station Abbreviations

The data dictionary would use data element abbreviations used in local databases such as Facilities Management or Military Manpower. For example:

- ACRN – Accounting Classification Reference Number
- AMD – Active Military Document
- CLIN – Contract Line Item Number
- DO – Delivery Order

Common Abbreviations

Likewise, commonly used abbreviations from everyday life should also be used whenever possible, such as:

- AVG – Average
- BLVD – Boulevard
- CORP – Corporation
- DEPT – Department

Proposed Shore Station Abbreviation Methodology

The staff recommends the following rules and methods to abbreviate attribute names when such abbreviations can not be found in the sources mentioned above. In general:

- Do not abbreviate colors due to the ambiguity that might arise. Consider blue and black.
- Do not abbreviate words five (5) characters or less. For words greater than five (5) characters, use the rules/guidelines listed below to abbreviate each word.
- Use the first four to six letters of the "root" word before beginning the shortening/discarding process. The "root" abbreviation should not exceed six (6) or seven (7) characters.
- Establish the "root" word abbreviation, then establish its family. For example, if the word being submitted is "CERTIFICATION", then using CERTIFY as the root (or parent), the family tree would look like:
    - The "root" CERTIFY would be abbreviated to CERTF (see rule 2)

- The “family” would then look like:
  - CERTFD (for certified)
  - CERTFG (for certifying)
  - CERTFR (for certifier)
  - CERTFN (for certification)

Abbreviation Caveats

- Although our desire is that every abbreviation be assigned to one and only one word, we recognize from the outset that that will not always be possible. Therefore, each abbreviation will have to be considered in the larger context of the application and the table within which it resides.
- The guidelines proposed herein are just that – guidelines – which may be broken with good reason from time to time by those acting as the data librarian, such as for clarity or to illustrate a particular business peculiarity.
- Some words just seem to defy an organized means of abbreviating them. In such cases, the data librarian, database administration staff, and application staff will strive to reach an agreeable compromise.
- Rules like we’ve defined below can be obscure. “Two vowels followed by a consonant followed by another vowel” sounds like we’re back in grade school. Look at the examples we’ve provided before throwing your hands up in disgust.
- Some of the rules overlap. When this occurs, a decision will be made by the data librarian and/or the database administration staff as to which rule to apply.
- At this time, the conventions proposed herein are for text-only data types.

Abbreviation Rules

Rule 1

If the word ends with one or more vowels followed by a consonant, drop the vowels, but keep the consonant unless it’s the same as the remaining one.

- **ail/al/an/ar/ed/en/er/ex/ial/ir/or/ual**

Examples:

DETAIL	becomes	DETL
INITIAL	becomes	INITL
ACTUAL	becomes	ACTL

Rule 2

If the word ends with one or more vowels followed by a consonant followed by a vowel, drop the vowels, but keep the consonant unless it’s the same as the remaining one.

- **age/uage/egy/ify/ive**

Examples:

LANGUAGE	becomes	LANG
STRATEGY	becomes	STRATG
ACTIVE	becomes	ACTV

Rule 3

If the word ends with one or more vowels followed by “TION” or “SION”, keep the “root” portion of the word (up to six characters) but replace the remaining string by an “N”, unless it’s the same as the remaining consonant.

- **tion/sion/iation/ation/ition/asion/iasion/ision**

Examples:

ACTION	becomes	ACTN
EXTENSION	becomes	EXTEN
ABBREVIATION	becomes	ABBREVN
DURATION	becomes	DURN

Rule 4

Similar to Rule 3, if the word ends with one or more vowels, one or more consonants, a vowel, followed by “TION” or “SION”, keep the “root” portion of the word (up to six characters) but replace the remaining string by an “N”, unless it’s the same as the remaining consonant.

- **tion/sion**

Examples:

CERTIFICATION	becomes	CERTFN
JUSTIFICATION	becomes	JUSTFN
VERIFICATION	becomes	VERFN

Rule 5

If the word ends with “ING”, keep the “root” portion of the word (up to six characters) but replace the remaining string by a “G”, unless it’s the same as the remaining consonant.

- **ing**

Examples:

ACCOUNTING	becomes	ACCTG
ENDING	becomes	ENDG
EXISTING	becomes	EXISTG

Rule 6

If the word ends with one or more vowels followed by “GHT”, drop the vowels and the “GH”.

- **eight**

Examples:

HEIGHT	becomes	HT
WEIGHT	becomes	WT

Discarding Rules

Some word spellings lend themselves to a discard process rather than the shortening process mentioned above. The following rules/guidelines are suggested to assist in that process.

Rule 7

If the word ends with one or more vowels followed by one or more consonants and a vowel, drop everything from the first vowel.

- **ace/ance/ease/ence/ency/ense/iance/iate/ily/ine/uance/ure/ute**

Examples:

INTERFACE	becomes	INTERF
CLEARANCE	becomes	CLEAR
CONFERENCE	becomes	CONFER

Rule 8

If the word ends with one or more vowels followed by one or more consonants and a “T”, drop everything from the first vowel.

- **act/aint/ect/ent/est/inct/ist/ort/uest/ust**

Examples:

• CONNECT	becomes	CONN
• RESIDENT	becomes	RESID
• REQUEST	becomes	REQUEST

Rule 9

If the word ends with one or more vowels followed by a “MATE”, “NATE”, “TIVE”, or “SIVE”, drop everything from the first vowel.

- **ative/itive/imate/sive/tive**

Examples:

• CUMULATIVE	becomes	CUMUL
• ESTIMATE	becomes	EST

Rule 10

If the word ends with one or more vowels followed by a “CAL”, drop everything from the first vowel.

- **ical**

Examples:

• CRITICAL	becomes	CRIT
• PHYSICAL	becomes	PHYS
• TECHNICAL	becomes	TECHN

Rule 11

If the word ends with one or more vowels followed by a “MENT” or “NENT”, drop everything from the first vowel.

- **ement/ment/anent/ument**

Examples:

• STATEMENT	becomes	STAT
• EQUIPMENT	becomes	EQUIP
• DOCUMENT	becomes	DOC

Conclusion

In its current iteration, the Enterprise Data Repository solves the three problems mentioned at the beginning of this report by:

- Housing the metadata from all Oracle databases regardless of operating system.
- Supporting data retrieval by ICBM functionality.
- Providing an organized means for data migration.
- Establishing standardized table and column conventions.

However, it has two major weaknesses in that:

(1) it can’t retrieve metadata from non-Oracle databases of which there’s an abundance; and

(2) it only supports text-type fields. It does not support audio, video, photographs, or spatial data.